DOCUMENT RESUME

ED 209 906 EC 140 626

AUTHOR Garrison, Wayne M.; Coggiola, Deborah C.
TITLE Practical Procedures for Test Length Reduction and Item Selection.
INSTITUTION National Technical Inst. for the Deaf, Rochester, N. Y.
SPONS AGENCY Department of Health, Education, and Welfare, Washington, D.C.
REPORT NO NTID-PS-38
PUB DATE Sep 80
NOTE 20p.; For related documents, see EC 140 611-630.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Abstract Reasoning; *Cognitive Tests; College Students; *Deafness; Higher Education; Mathematics; Postsecondary Education; Reading Tests; Spatial Ability; *Student Evaluation; *Test Construction; Test Reliability; Test Validity

ABSTRACT

The study involving approximately 300 deaf students entering the National Technical Institute for the Deaf was designed to determine the extent to which four cognitive measures could be reduced in overall length without serious threat of information loss. The measures investigated were the California Reading Comprehension Test, the Mathematical Skills Program, the Abstract Reasoning Test, and the Space Relations Test. In all instances, the correlation between reduced and full length tests is very high and positive, indicating that control over the distribution of the difficulty indices associated with the items in a test produces efficient tests requiring substantially fewer stimuli than those included in the original design sets. Moreover, the effects of test stimuli reduction on recomputations of the reliability estimate for each test reveals negligible decreases. In summary, the findings showed that the number of test stimuli comprising the four instruments studied could be reduced by an average of 54%. (SB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED209906

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

PRACTICAL PROCEDURES FOR TEST LENGTH REDUCTION AND ITEM SELECTION

Paper Series No. 38
September, 1980

Wayne M. Garrison
Deborah C. Coggiola

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
NTID
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

NTID's principal goal in doing research is to influence the education, training and career placement of deaf citizens through systematic examination of issues related to deafness. As one part of NTID's total research effort, the Department of Research and Development conducts descriptive and experimental research. Research findings are used in the development of programs and materials in the areas of learning and instruction, personal and social growth, and career development of deaf students. This document was developed in the course of an agreement with the U.S. Department of Health, Education and Welfare.

EC 140626

Abstract

This study investigated the effects of test item eliminations on the prediction accuracy of examinees' total test scores for each of four different cognitive skill measures. Instruments subjected to analysis were the reading subscale of the California Achievement Test, the Space Relations and Abstract Reasoning subtests of the Differential Aptitude Test battery, and a test of mathematical skills developed at the National Technical Institute for the Deaf (NTID). The respondents consisted of groups of incoming deaf students tested during the 1979 and 1980 NTID summer orientation programs. When the items comprising the four instruments were examined in terms of their difficulty and, hence, fit to the target populations, it was found that the instruments could be reduced in length by an average of 54% of the original number of items. The mean correlation between the reduced and full-length test scores was found to be .96, indicating negligible information loss with substantially shortened measures.

Practical Procedures for Test Length Reduction and Item Selection

Between the years 1973 and 1978, a comprehensive battery of cognitive and communication skill measures was administered annually to students attending NTID's Summer Vestibule Program (SVP). In all, the battery included more than twenty separate tests, requiring some 25-30 hours in administration, scoring, and interpretation (for a description of the tests, see Long & Coggiola, 1980). Additionally, students were sometimes tested by individual academic departments for secondary admission, placement and/or advisement purposes (Coggiola, Long & Streim, 1980).

Recognizing the demands placed on students' time by such an extensive testing program, serious questions were raised regarding its continuation. Specifically, it was asked whether all of the tests included in the battery provided information useful in the prediction of students' academic performance at NTID. That is, could the battery be reduced to fewer components? If so, which measures could be eliminated? Most importantly, what level of performance predictability could be demonstrated for those measures which were retained in the battery?

White (1979) initiated the evaluation of the testing program at NTID. Using data obtained from administration of the 21 measures of cognitive and communication skill included in the overall test battery, White utilized the methodology of multiple regression analysis to explore the utility of the instruments in predicting student achievement at NTID. Using cumulative grade point average (GPA) as the criterion variable White found, that across three different study samples, only five of the 21 measures analyzed contributed significantly to the explainable variance in GPA. As a result of this study, testing

activities during SVP 1979 and SVP 1980 were limited to assessments on only five of the 21 tests comprising the battery.

Long and Coggiola (1980) also examined the testing program at NTID. They, like White (1979), sought to identify a subset of measures from the extensive test battery which would provide the most non-redundant information about students' academic performance at NTID. Long and Coggiola utilized a larger sample of subjects than did White, and exercised additional controls over the data structures. Specifically, students with fewer than 30 accumulated credit hours of course work were excluded from analyses, and regressions were performed on students' data grouped according to particular career areas (five total) and specific programs of study (19 total).

Long and Coggiola (1980) found that White's (1979) results generally were substantiated for the larger sample of individuals studied. However, the relative importance of each of the five measures included in the regression analyses varied across career areas and academic programs. This was so because academic programs of study differ in both the kind and proficiency of skills required for subject matter mastery.

Having reduced the larger test battery to a subset of five relatively independent measures of cognitive functioning, Garrison and Long (1980) sought to determine whether those measures which utilized an objective response format (i.e., right/wrong item scoring) could be fitted to a mathematical measurement model. Four of the five cognitive skill measures were relevant to the analysis. The NTID test of writing ability was not subjected to fit analysis insofar as proficiency on this measure involves scorer subjectivity.

The analysis of fit of a set of test data to a measurement model is essentially a hypothesis-testing activity, whereby one postulates, and then evaluates, the operation of certain (item) factors in examinees' observed test performance. The particular functional

relationship between test items and the ability continuum varies among the models (e.g., polynomials, cumulative normal ogive, logistic curve). However, all of the existing models proposed for person measurement conceptualize an individual's success or failure on any test item in terms of: (1) an examinee's position on an ability continuum, and (2) one or more constants associated with each of the items comprising a test.

Using data obtained from 288-308 students tested during SVP 1979, Garrison and Long (1980) examined the fit of each of the four cognitive skill measures (referred to above) to the Rasch measurement model. Within the limits of statistical probability, model/data fit was achieved in all instances, following specific item and person eliminations from the initial data matrices. This finding provides an additional source of evidence that the measures are usable with the NTID deaf student population. Moreover, analyses of the data showed that examinees' inferred ability levels remained relatively invariant across randomly generated subsets of items from each test, illustrating the item-free person measurement aspect of Rasch test calibration. [Note, however, that randomly generated test forms are not necessarily maximally efficient tests.]

In summary, the findings of White (1979), and those of Long and Coggiola (1980), indicated that a descriptive summary of students' cognitive skills and academic performance could be derived using the five measures which were subjected to in-depth analyses. These five measures are intended to assess individuals' acquisition of basic reading, writing, and mathematical skills, and abstract reasoning and spatial abilities. Moreover, the findings of Garrison and Long (1980) indicated that assessments obtained with four of these measures conformed to the theoretical expectations of the Rasch measurement model. That a set of test data may be found to fit the Rasch model implies that the variable measured is unidimensional, and that items and persons do not differ substantially with respect to response factors not represented in the model. That is, the outcome of

person/item interactions in a testing situation is explainable simply in terms of the difference between an examinee's ability on some latent continuum and the difficulty of items scaled along the same continuum.

In the same manner that multiple regression techniques may be used to "simplify" an initial set of independent or predictor variables through the elimination of those variables which do not add appreciably to prediction accuracy, so do item analyses serve the useful purpose of identifying those stimuli in a test which do not contribute substantially to measurement efficiency. With the aim of extending the research summarized in the preceding paragraph, the present study was undertaken to determine the extent to which each of the four measures analyzed earlier by Garrison and Long (1980) could be reduced in overall length, without serious threat of information loss. Specifically, the purpose of the present study was to examine the effects of test item eliminations on the prediction accuracy of examinees' total test scores.

METHOD

Data Source

The primary sample of respondents in this study consisted of 288-308 male and female deaf students entering NTID during the fall, 1979. Data were obtained as part of a summer sampling program offered immediately prior to formal entrance into NTID. A secondary sample of 299-309 deaf students entering NTID during the fall, 1980 was used for cross-validation of the statistical results.

Instrumentation

The cognitive skill measures subjected to analysis in this study represented measures of reading "comprehension", mathematical abilities, abstract reasoning, and spatial abilities. In all instances, the tests were administered in group settings over a period of two to three days. Each of the instruments is described briefly below:

(1) California Reading Comprehension Test (CAT-R: Tiegs & Clark, 1963)

The CAT-R is a subtest of the larger California Achievement Test series and consists of items relating to following directions, reference skills and the interpretation of text material. The test is comprised of 86 items total which, collectively, represent a measure of verbal skills. The CAT-R is used for the initial placement of deaf students into the NTID English program. [Later reference to this test will use the shortened abbreviation CAT.]

(2) Mathematical Skills Program (MSP: Taylor, 1979)

The MSP was developed at NTID in the Math Learning Center. It is comprised of 60 items measuring basic abilities in the use and manipulation of whole numbers, fractions, decimals and percentages. The MSP is used for the placement of students into the NTID Mathematics program.

(3) Abstract Reasoning Test (ABT: Bennett, Seashore & Wesman, 1966)

The ABT is intended to be a nonverbal measure of an individual's reasoning ability and is part of the larger Differential Aptitude Test battery. The test requires the perception of an operating principle in a series of changing diagrams. It consists of 50 items total and is administered under timed conditions. The test is used as part of the career counseling process with NTID students.

(4) Space Relations Test (SRE: Bennett, Seashore & Wesman, 1966)

The SRE is also a subtest of the Differential Aptitude Tests and is intended to be a measure of an individual's ability to visualize objects in picture form and to imagine how the objects would appear if rotated in various ways. The test manual suggests that the SRE is a measure of the ability to deal with concrete

material through visualization. The test consists of 60 items and is administered under strict time limits. Like the ABT, scores on this test are used in career counseling of NTID students. Both the SRE and ABT have been found to have greatest predictive validity in art related areas of study at NTID (Long & Coggiola, 1980).

Analysis

Data obtained from each of the instruments described in the previous section were subjected to Rasch calibration using program BICAL (Wright & Mead, 1977). The calibration of instruments with this program results in the estimation of item difficulties and person abilities expressed in the same units of measurement (i.e., logits). These estimates then provide all of the information necessary to relate the characteristics of tests to the characteristics of the target population. The mathematical details of the estimation process are presented elsewhere (e.g., Wright & Stone, 1979). Briefly, parameter estimation is accomplished through the method of unconditional maximum likelihood, the estimated values of the item difficulty and person ability parameters being such that the actual pattern of responses observed in the data (a set of correct/incorrect responses by items and persons) is the most likely.

RESULTS AND DISCUSSION

While a variety of statistical indices are used in the course of test analyses to describe the individual items comprising an instrument, the item difficulty index is the most straightforward. When examinees' test scores are evaluated in the light of the item difficulty parameter, it is possible to make probabilistic statements about any particular examinee's success/failure on any test item. Moreover, over a sufficiently large number of test cases, the item difficulty information may be used to determine the characteristics of a "best" test for a particular population of respondents with whom the test will be used in the future.

Figure 1 illustrates the distribution of scores (expressed in the Rasch metric) of the 308 persons tested with the MSP during 1979. This pictorial representation of the group's performance on the MSP may be reduced to a statistical statement containing three parameters. These parameters correspond to the target's center (mean), dispersion (standard deviation), and approximate shape (e.g., normal, interval).

When the items in a test operate in ways that fit the Rasch model, then the operating curve for a test also lends itself to specification in terms of only three parameters. These three parameters correspond to the average difficulty of the items comprising a test, the range of item difficulties, and the number of items included in a test. Wright and Stone (1979) appropriately termed these three test characteristics height, width and length, respectively.

It is axiomatic in measurement theory that a "best" test is that which maximizes measurement in the region in which measurement is expected to occur. Consequently, the design of a best test depends on relating the characteristics of the target (i.e., center, dispersion, distribution shape) to the characteristics of a test (i.e., height, width, length). This matching of test to target for the 1979 MSP data was accomplished in the following manner:

1. Calibration of the MSP revealed a mean ability (M) of .32 logits for the 308 students tested, with standard deviation (S) equal to 1.18 (see Figure 1). The mean difficulty level (H) of the items was .00 logits with standard deviation (W) equal to 1.13, indicating that the MSP, in general, was less difficult than the examinees, as a group, were able.

2. From the distribution of observed scores on the MSP, the mean difficulty level of the best fitting test for this target group of respondents was defined as H = M.

3. With an approximate normal distribution (N) of examinee ability on the MSP, and S = 1.18, the value of W suggested by Wright and Stone (1979, pg. 139) for the most efficient uniform test is one logit.

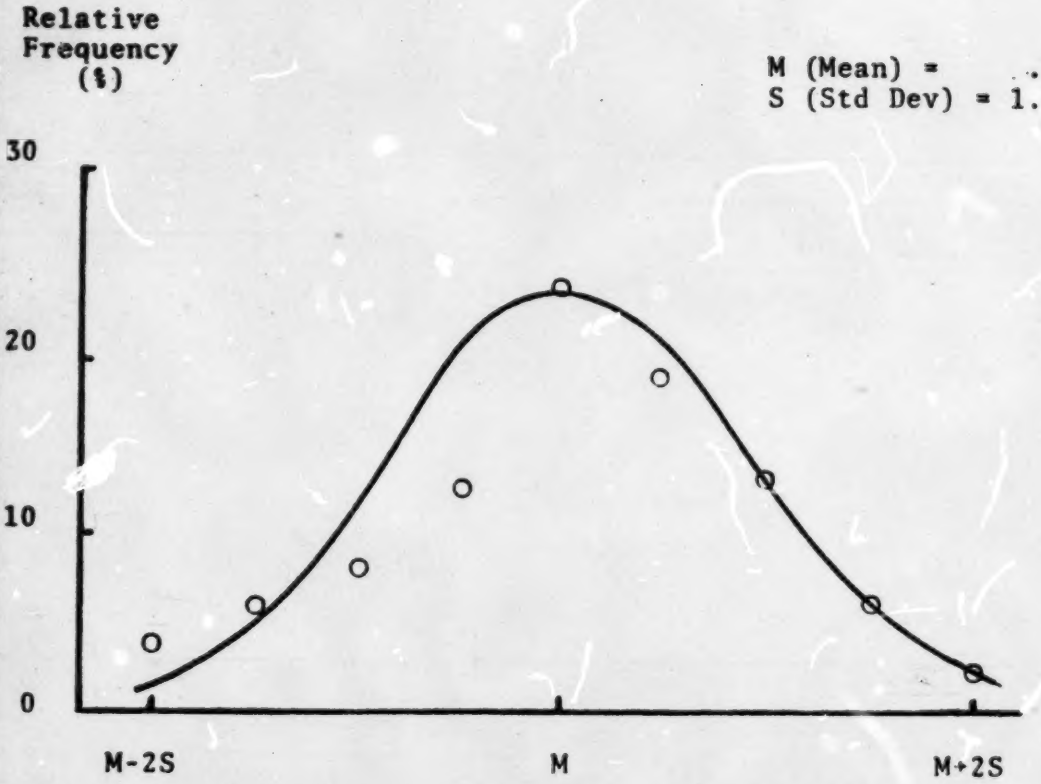

Figure 1. MSP score distribution.

4. With optimal W = 1, it can be shown that a test composed of items with difficulty values in the range H ± W/2 minimizes error of measurement over a target of examinees characterized by $N(M, S^2)$ [See Wright and Stone, 1979].

5. From the pool of 60 MSP items, a refined test was formed. This test was composed of items with difficulty estimates in the range .32 ± .50 logit. The net effect of this target/test matching process is the elimination of items either too easy or too difficult for the population under investigation. Of course, items contained in the refined test version which failed to meet the statistical criteria of fit to the Rasch model (i.e., unstable difficulty estimates across differing levels of examinee ability) were also eliminated. [For a discussion of item fit to the Rasch model, see Garrison and Long, 1980.]

6. From steps 1-5 above, the original 60-item MSP was reduced to a refined version containing 25 items.

To determine how well this new subset of MSP items conveyed information about the respondents' levels of ability, relative to the information provided by the full-length test, several correlational analyses were performed. The correlation between respondents' test scores on the 25-item and 60-item tests was found to be .96. The internal consistency reliability estimate for the refined test compared favorably with that obtained from analysis of the full-length test (.89 and .93, respectively).[1] Finally, when the data obtained from administration of the MSP during 1980 were examined for cross-validation of the 1979 results (i.e., different examinees, identical part-whole test comparison), the subtest-total test correlation was also found to be .96. Internal consistency reliability estimates for the 60 and 25-item tests (1980 data) were found to be .93 and .88, respectively.

In the same manner that analyses were performed on the MSP data, statistical analyses were performed on the other three measures included in this study. Following the calibration of each of the instruments with the Rasch scaling procedure, subsets of items were identified with average difficulty (H) set at the mean of the observed score distributions, with new dispersion (W) determined by tabled values provided by Wright and

Stone (1979). From these potentially optimal tests, items which failed to conform to the statistical expectations of the Rasch model were subsequently eliminated.

Table 1 summarizes the results of analyses performed on all tests from data gathered during 1979. In all instances, the correlation between reduced and full-length tests is very high and positive, indicating that control over the distribution of the difficulty indices associated with the items in a test produces efficient tests requiring substantially fewer stimuli than those included in the original design sets. Moreover, the effects of test stimuli reduction on re-computations of the reliability estimate for each test reveals negligible decreases.

When the data obtained from the 1980 test administrations were analyzed for cross-validation of the 1979 results, the findings were remarkably similar. That is, using the item sets generated from the 1979 data analyses, the part-whole test correlations computed from the 1980 data were identical to those obtained from analysis of the 1979 data. Moreover, the reliability analyses of the separate data sets were comparable. Table 2 summarizes the results of analyses performed on the 1980 data. Items retained in each of the reduced tests are identified in the appendix of this report.

CONCLUSIONS

When a measurement of some characteristic, attribute or ability is planned, it is done so to acquire information about a particular target person or group of persons. The quality of the proposed measurements is dependent upon the extent to which instruments are constructed with the specifics of the target in mind. The design of an efficient test depends upon relating the characteristics of the target to those of the test stimuli, so that an optimal information gain/testing time condition is achieved.

Table 1
Summary 1979 Data Reduction Analyses

Test	Initial Test Length	Reduced Test Length	% Reduction	Subtest-Total Correlation	Alpha[1] Full Length	Alpha Reduced Length
MSP (n=308)	60	25	58.3	.96	.93	.89
CAT (n=303)	86	42	51.2	.97	.90	.86
SRE (n=289)	60	22	63.3	.94	.93	.88
ABT (n=288)	50	29	42.0	.96	.94	.89

[1]Coefficient alpha is a maximum likelihood estimate of test reliability, determined from a single administration of a test.

-11-

Table 2
1980 Cross-Validation Analyses

Test	Sample Size	Subtest-Total Correlation	Alpha Full Length	Alpha Reduced Length
MSP	309	.96	.93	.88
CAT	299	.97	.89	.85
SRI	306	.94	.93	.88
ABT	309	.96	.93	.87

The present study demonstrates that, when the items in a test operate in a manner that fits the Rasch measurement model, then the selection of an efficient set of calibrated items from a larger item pool requires attention only to the item difficulty parameter. When the operationalization of a test is based upon the difficulties of its items, it is possible to minimize the number of test tasks presented to a group of examinees, as observed in the present study, and, yet, maintain the quality of information afforded by more lengthy instruments.

In summary, the findings of this study indicate that the number of test stimuli comprising the four instruments studied could be reduced by an average of 54 percent. The correlation between scores on the reduced measures and the full-length instruments averaged .96 for the 1979 data. This correlational finding was upheld in cross-validation on the 1980 NTID testing data. As a cautionary note, however, it should be borne in mind that any systematic control over the properties of the items comprising a reduced form of a test should take into consideration the purpose served by the original test (e.g., prescriptive versus predictive). Moreover, the reduced form must preserve the balance of content and process characteristic of the original design set of items. The final selection of items, therefore, must be a compromise of several considerations.

REFERENCES

Bennett, G. K., Seashore, H. G., & Wesman, A. G. Differential Aptitude Tests. New York: Psychological Corporation, 1966.

Coggiola, D., Long, G., & Streim, N. Test use and interpretation by fourteen departments at NTID. Unpublished manuscript, National Technical Institute for the Deaf, 1980.

Garrison, W. M., & Long, G. L. Applications of the Rasch measurement model to cognitive assessments of deaf individuals. Paper presented at the Fourth International Symposium on Educational Testing, Antwerp, Belgium, 1980.

Long, G. L., & Coggiola, D. Prediction of academic performance and student classification using five cognitive skill measures. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Taylor, R. W. W. MSP: The math skills project at the National Technical Institute for the Deaf. Report No. 20, Technical Mathematics Department, National Technical Institute for the Deaf, 1979.

Tiegs, E. W., & Clark, W. W. California Reading Tests. San Diego: California Testing Bureau, 1963.

White, K. R. An evaluation of the testing program at the National Technical Institute for the Deaf. Unpublished manuscript, National Technical Institute for the Deaf, 1979.

Wright, B. D., & Mead, R. J. BICAL: Calibrating rating scales with the Rasch model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago, 1977.

Wright, B. D., & Stone, M. H. Best test design: Rasch measurement. Chicago: Mesa Press, 1979.

Footnote

[1]In general, the reliability or internal equivalence of a test is a function of both the item intercorrelations and the number of items. Other things being equal, reducing the length of a test can be expected to lower the magnitude of a reliability coefficient, as observed in the present study. However, the average item intercorrelations ($\bar{r}_j$, j = 1, 4) for the reduced tests generally were higher than those obtained for the full-length instruments, indicating improved internal consistency.

APPENDIX
Items Retained in Reduced Measures

Test	Item Numbers
MSP (1980 Form)	2, 7, 10, 16, 18, 20-22, 24-25, 27-28, 30, 32, 34, 40-41, 45, 47-48, 50, 55-58
CAT (Form W)	2, 9, 12-14, 19-21, 28, 30-34, 36, 38-41, 43-44, 50, 55-56, 59, 61-63, 65-66, 68-71, 72-73, 75-80
SRE (Form L)	29-50
ABT (Form L)	21-27, 29-50